Journal of the Korean Statistical Society (2023) 52:1055–1077
https://doi.org/10.1007/s42952-023-00233-4

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

A nonparametric binomial likelihood approach for causal inference in instrumental variable models

Kwonsang Lee[1] · **Bhaswar B. Bhattacharya**[2] · **Jing Qin**[3] · **Dylan S. Small**[2]

Received: 25 January 2023 / Accepted: 13 September 2023 / Published online: 19 October 2023

Abstract

Instrumental variable methods allow for inference about the treatment effect by controlling for unmeasured confounding in randomized experiments with noncompliance. However, many studies do not consider the observed compliance behavior in the testing procedure, leading to loss of power. In this paper, we propose a novel nonparametric likelihood approach, referred to as the *binomial likelihood* method, that incorporates information on compliance behavior while overcoming several limitations of previous techniques. Our proposed method produces proper estimates of the counterfactual distribution functions by maximizing the binomial likelihood over the space of distribution functions. Using this we propose two versions of a *binomial likelihood ratio test* for the null hypothesis of no treatment effect, and study their finite sample and asymptotic properties. We also develop an efficient algorithm for computing our estimates, and apply the method to study the effect of Medicaid coverage on mental health using the Oregon Health Insurance Experiment.

Keywords Causal inference · Distributional treatment effect · Nonparametric likelihood · Nonparametric two-sample test

✉ Kwonsang Lee
kwonsanglee@snu.ac.kr

Bhaswar B. Bhattacharya
bhaswar@wharton.upenn.edu

Jing Qin
jingqin@niaid.nih.gov

Dylan S. Small
dsmall@wharton.upenn.edu

[1] Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[2] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, 265 South 37th Street, Philadelphia, PA 19104, USA

[3] Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, 5601 Fishers Lane, Rockville, MD 20892, USA

1 Introduction

The instrumental variables (IV) method is a popular technique for estimating the casual effect of a treatment in the presence of unmeasured confounding (Angrist et al., 1996; Baiocchi et al., 2014; Tan, 2006). This arises in situations where, even through direct randomization is impossible, an encouragement to take the treatment can be randomized (Holland, 1988), or there is a "natural experiment" such that some people are encouraged to receive the treatment compared to others in a way that is effectively random (Angrist et al., 1996). Informally, an instrument is a variable that affects the treatment but is independent of unmeasured confounders and only affects the outcome through affecting the treatment (see Sect. 2.1 for a more precise definition). Under a monotonicity assumption that the encouraging level of the instrument never causes someone not to take the treatment, the treatment effect can be identified for the compliers, those subjects who would take the treatment if they were encouraged to take the treatment but would not take the treatment if they were not encouraged [see Angrist et al. (1996), Abadie (2003), Baiocchi et al. (2014), Brookhart and Schneeweiss (2007), Cheng et al. (2009a, 2009b), Hernan and Robins (2006), Kang et al. (2018), Johnson et al. (2019), Ogburn et al. (2015), Tan (2006) and the references therein for methods of inference using instrumental variables].

In causal inference, to evaluate the treatment effect on the outcome, Fisher's sharp hypothesis of no effect is often considered, which, in the potential outcome framework (Neyman, 1923; Rubin, 1974), asserts that the two potential outcomes $Y_i(1)$ and $Y_i(0)$, which are the outcomes individual $i \in \{1, 2, \ldots, n\}$ would experience with or without treatment, respectively, are the same for every individual i. Under the IV assumptions, where the treatment effect can be identified for the compliers, the hypothesis of no effect for compliers can be tested by comparing the distributions of $Y_i(1)$ and $Y_i(0)$ for compliers. Unfortunately, it is difficult to make inference about these distributions since researchers do not know who are the compliers from data. Abadie (2002) proposed an approach that indirectly compares the two potential outcome distributions, by using the Kolmogorov–Smirnov test statistic. However, this approach ignores the treatment variable during the testing procedure. Thus, it does not consider the compliance class information of the individuals, which can lead to loss of power, as discussed in Rubin (1998).

In this paper, we propose a novel nonparametric likelihood-based approach for comparing the two counterfactual distribution functions, with or without treatment for compliers, that uses the compliance class information and allows for estimation and hypothesis testing in a common holistic framework. This requires a methodological innovation because the usual nonparametric likelihood approach using the empirical likelihood (Owen, 2001) does not work for the IV model because there are infinitely many solutions that maximize the likelihood (Geman & Hwang, 1982). Our proposed *binomial likelihood* (BL) approach creates a piece of likelihood at each knot (or evaluating point), by using binomially distributed outcomes: outcomes smaller than or equal to the knot, and outcomes

larger than the knot. Then, it multiplies together the pieces of these likelihoods across all knots creating a composite likelihood. This is a "pseudo" likelihood rather than the true likelihood because the binomial random variables are actually dependent, but are treated as independent in the composite likelihood. Due to its binomial nature in defining likelihood functions, we specifically call this composite likelihood, the binomial likelihood (BL). Composite likelihood has been found useful in a range of areas including problems in geostatistics, spatial extremes, space-time models, clustered data, longitudinal data, time series and statistical genetics; see Lindsay (1988), Heagerty and Lele (1998), Larribe and Fearnhead (2011), and Varin et al. (2011).

The BL approach can be used for statistical inference similar to the usual likelihood method. For instance, for estimating the distribution functions of the compliers, the *maximum binomial likelihood (MBL) estimate* can be obtained by maximizing the BL over the space of distribution functions. Therefore, by definition, the MBL estimates satisfy the necessary conditions for a proper distribution function (increasing and non-negative). This make the BL estimates easily interpretable, and is a major improvement over the naive *plug-in* estimates, which can be non-monotonic and negative. As a consequence, the BL method can be effectively used for making further inferences, such as integrating utility functions or estimating moments of the probability function. Furthermore, similar to classical likelihood ratio tests, the *binomial likelihood ratio test (BLRT)* for the null hypothesis of no treatment effect can be constructed by taking the ratio of two BL values that are maximized over the null and the alternative respectively. For computing the MBL estimate and conducting hypothesis testing using the BLRT we develop a computationally efficient iterative algorithm based on the expectation-maximization (EM) and pool-adjacent-violators (PAV) algorithms. Thus, the BL approach provides the practitioners with a comprehensive toolbox for causal inference in non-parametric IV problems.

The BL method has several attractive limiting and finite-sample properties. To begin with, we show that the MBL estimate for the distribution function of the compliers has the same first-order asymptotics (limiting distribution) as the naive *plug-in* estimates. This shows that the BL estimates, which preserve all the properties of a proper distribution, have no loss in asymptotic efficiency compared to the naive estimates, which can be non-monotone and negative in finite samples. For hypothesis testing, we show that the BLRT is asymptotically equivalent to the well-known Anderson–Darling two-sample test (Pettitt, 1976). Since there are no closed form expressions for the BL estimates in general, these asymptotic results are important to the understanding of the BL approach. The BLRT also has better finite-sample performance for detecting distributional changes compared to other baseline methods. The improvement is especially significant in the weak IV setting, exhibiting the importance of incorporating the compliance class information for hypothesis testing in IV models. We also apply the BL approach to study the effect of Medicaid coverage for African American adults on self-reported mental health, as studied by Baicker et al. (2013).

The rest of the article is organized as follows. Basic notation and assumptions of the IV model are discussed in Sect. 2. In this section, we also review the existing

plug-in approach for testing the hypothesis of no effect. In Sect. 3, we introduce the BL approach and derive the asymptotic properties of the MBL estimate (Theorem 1). In Sect. 4, we develop two versions of the BLRT for testing the null hypothesis, and derive the asymptotic properties of the tests (Theorems 2 and 3). In Sect. 5 we discuss the algorithm for computing the BL estimates and present the numerical results for the BLRT. The analysis of the real data is given in Sect. 6. Proofs of the theorems and additional simulations are given in the supplementary materials.

2 Framework and review

2.1 Assumptions and identification with instrumental variables

For individual i, denote Z_i as the binary IV, D_i as the indicator variable for whether individual i receives the treatment or not, and Y_i as the outcome variable that is continuous in this paper. Using the potential outcome framework (Neyman, 1923; Rubin, 1974), define $D_i(0)$ as the value that D_i would be if Z_i were to be set to 0, and $D_i(1)$ as the value that D_i would be if Z_i were to be set to 1. Similarly, $Y_i(z,d)$ for $(z,d) \in \{(0,0),(0,1),(1,0),(1,1)\}$, is the value that the outcome Y_i would be if $Z_i = z$ and $D_i = d$. For each individual i, the analyst can only observe one of the two potential values $D_i(0)$ and $D_i(1)$, and one of the four potential values $Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)$. The observed treatment D_i is $D_i = Z_i D_i(1) + (1 - Z_i)D_i(0)$. Similarly, the observed outcome Y_i can be expressed as $Y_i = Z_i D_i Y_i(1,1) + Z_i(1 - D_i)Y_i(1,0) + (1 - Z_i)D_i Y_i(0,1) + (1 - Z_i)(1 - D_i)Y_i(0,0)$. An individual's *compliance class* is determined by the combination of the potential treatment values $D_i(0)$ and $D_i(1)$, which is denoted by S_i: $S_i = \textit{always-taker}$ (at) if $D_i(0) = D_i(1) = 1$; $S_i = \textit{never-taker}$ (nt) if $D_i(0) = D_i(1) = 0$; $S_i = \textit{complier}$ (co) if $D_i(0) = 0, D_i(1) = 1$; and $S_i = \textit{defier}$ (de) if $D_i(0) = 1, D_i(1) = 0$.

For the rest of this paper, the following standard identifying conditions are assumed. The implications of these conditions are briefly explained in the paragraph below; see Angrist et al. (1996) for more details on these conditions.

Assumption 1 The following identification conditions will be imposed on the instrumental variable model:

(a) Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1986): The outcome (treatment) for individual i is not affected by the values of the treatment or instrument (instrument) for other individuals and the outcome (treatment) does not depend on the way the treatment or instrument (instrument) is administered.
(b) The instrumental variable Z_i is independent of the potential outcomes $Y_i(z,d)$ and potential treatment $D_i(z)$.

$$Z_i \perp\!\!\!\perp \big(Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), D_i(0), D_i(1)\big)$$

(c) Nonzero average causal effect of Z_i on D_i: $\mathbb{P}(D_i(1) = 1) > \mathbb{P}(D_i(0) = 1)$.
(d) *Monotonicity* $D_i(1) \geq D_i(0)$.

(e) *Exclusion restriction* $Y_i(0,d) = Y_i(1,d)$, for $d = 0$ or 1.

Assumption 1 enables the causal effect of the treatment for the subpopulation of the compliers to be identified. Condition (a) allows us to use the notation $Y_i(z,d)$ (or $D_i(z)$), which means that the outcome (treatment) for individual i is not affected by the values of the treatment and instrument (instrument) for other individuals. Condition (b) will be satisfied if Z_i is randomized. Condition (c) requires Z_i to have some effect on the average probability of treatment. Condition (d), the monotonicity assumption, means that the possibility of $D_i(0) = 1$, $D_i(1) = 0$ is excluded, that is, there are no defiers. Condition (e) assures that any effect of Z_i on Y_i must be through an effect of Z_i on D_i. Under this assumption, the potential outcome can be written as $Y_i(d)$, instead of $Y_i(z,d)$.

Let $\phi_1 = \mathbb{P}(Z = 1)$, $\phi_s = \mathbb{P}(S = s), s \in \{co, nt, at\}$. Also, let $F_{co}^{(0)}(t), F_{nt}(t), F_{co}^{(1)}(t)$, and $F_{at}(t)$ be the cumulative distribution functions of the outcome Y for compliers without treatment, never-takers, compliers with treatment, and always-takers respectively. For $F_{co}^{(0)}(t)$ and $F_{co}^{(1)}(t)$, under Assumption 1, they are identified as the distributions of the potential outcome $Y(0)$ and $Y(1)$ respectively, for example, $F_{co}^{(0)}(t) = \mathbb{P}(Y(0) \le t \mid S = co)$. Similarly, we define the distribution functions of Y corresponding to combinations of Z, D. Denote $F_{zd}(t) = \mathbb{P}(Y \le t \mid Z = z, D = d)$. Although $F_{Y|zd}$ can be more accurate notation than F_{zd} since Z and D are conditioned on, we will instead use simpler notation F_{zd}. Any notation involving F followed by a subscript means the distribution function of Y conditioning on the subscript. Also, we define the probabilities $\eta_{zd} = \mathbb{P}(Z = z, D = d)$ for $z, d \in \{0, 1\}$. Finally, let $H(t) = P(Y \le t) = \sum_{z,d\in\{0,1\}} \eta_{zd} F_{zd}(t)$, be the mixture distribution of F_{zd}. The outcomes $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed from $H(t)$. Under Assumption 1, as discussed in Abadie (2002), both $F_{co}^{(0)}(t)$ and $F_{co}^{(1)}(t)$ can be identified as

$$F_{co}^{(0)}(t) = \frac{(\phi_{co} + \phi_{nt})F_{00}(t) - \phi_{nt}F_{10}(t)}{\phi_{co}}, \quad F_{co}^{(1)}(t) = \frac{(\phi_{co} + \phi_{at})F_{11}(t) - \phi_{at}F_{01}(t)}{\phi_{co}}. \tag{1}$$

Also, $F_{nt}(t)$ and $F_{at}(t)$ can be identified under Assumption 1 as $F_{nt}(t) = F_{10}(t)$ and $F_{at}(t) = F_{01}(t)$.

2.2 Testing Fisher's null hypothesis of no effect: review of the existing approaches

A central question in causal inference is to understand if the treatment has any causal effect on the outcome. To evaluate the treatment effect on the outcome, Fisher's sharp hypothesis of no effect can be considered. Under Assumption 1, it can be tested whether there is any causal treatment effect for compliers. Technically, Fisher's hypothesis can be constructed for compliers as $H_0^{\text{compliers}} : Y_i(1) = Y_i(0)$ for $S_i = co$. However, $H_0^{\text{compliers}}$ cannot be directly tested since only one of the two potential outcomes for each individual can be observed. Instead, we consider a test for equality of distributions using the potential outcome distributions for compliers,

$$H_0 : F_{co}^{(0)}(t) = F_{co}^{(1)}(t), \text{ for all } t \in \mathbb{R}. \tag{2}$$

The existing approach for testing H_0 is based on the fact that $F_{co}^{(0)}(t) = F_{co}^{(1)}(t)$ implies $F_0(t) = F_1(t)$ where $F_z(t) = \mathbb{P}(Y \le t \mid Z = z)$ under Assumption 1. Abadie (2002) proposed using the Kolmogorov–Smirnov test $T_{KS} = \sup_{t\in\mathbb{R}} |\overline{F}_0(t) - \overline{F}_1(t)|$, where $\overline{F}_z(t) = \sum_{i=1}^n \mathbf{1}\{Y_i \le t, Z_i = z\}/\sum_{i=1}^n \mathbf{1}\{Z_i = z\}$ are the empirical distribution functions for $z = 0, 1$. This test is the comparison between the outcome distribution of the $Z = 0$ group and the outcome distribution of the $Z = 1$ group. To show a connection between these two distributions with the compliers' distribution functions $F_{co}^{(0)}(t)$ and $F_{co}^{(1)}(t)$, define the *plug-in* estimates obtained using (1) as

$$\breve{F}_{co}^{(0)}(t) = \frac{\left(\breve{\phi}_{co} + \breve{\phi}_{nt}\right)\overline{F}_{00}(t) - \breve{\phi}_{nt}\overline{F}_{10}(t)}{\breve{\phi}_{co}}, \ \breve{F}_{co}^{(1)}(t) = \frac{\left(\breve{\phi}_{co} + \breve{\phi}_{at}\right)\overline{F}_{11}(t) - \breve{\phi}_{at}\overline{F}_{01}(t)}{\breve{\phi}_{co}}, \tag{3}$$

where $n_{zd} = \sum_{i=1}^n \mathbf{1}\{Z_i = z, D_i = d\}$, $\breve{\phi}_{nt} = n_{10}/(n_{10} + n_{11})$, $\breve{\phi}_{at} = n_{01}/(n_{00} + n_{01})$, $\breve{\phi}_{co} = 1 - \breve{\phi}_{nt} - \breve{\phi}_{at}$, and $\{\overline{F}_{zd}\}_{z,d\in\{0,1\}}$ are the empirical distribution functions, $\overline{F}_{zd}(t) = (1/n_{zd}) \sum_{i=1}^n \mathbf{1}\{Y_i \le t, Z_i = z, D_i = d\}$. Since $|\overline{F}_0(t) - \overline{F}_1(t)| = |(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)) \cdot \breve{\phi}_{co}|$, T_{KS} is equivalent to the test based on comparison between $\breve{F}_{co}^{(0)}$ and $\breve{F}_{co}^{(1)}$. However, the plug-in estimates have two limitations: (1) violating the non-decreasing condition of distribution functions and (2) being unstable when an IV is weak. First, the violation leads to producing estimates that are often located outside of [0, 1]. Therefore, they are not proper estimates of $F_{co}^{(0)}$ and $F_{co}^{(1)}$, which is due to not incorporating the observed information on compliance behavior. Furthermore, the test statistic T_{KS} can be misleading since it is based on proper distribution functions $\bar{F}_0$ and $\bar{F}_1$ even when the fluctuations of $\breve{F}_{co}^{(0)}$ and $\breve{F}_{co}^{(1)}$ are severe. This issue often occurs when an IV is weak. As discussed in Rubin (1998), making use of the IV structure can produce a better test statistic, and thus increase power.

To employ the structure of the instrumental variable model, one simple way is to transform the plug-in estimates to proper distribution functions by using the monotone rearrangement method (Chernozhukov et al., 2010), followed by truncation to [0, 1]. Chernozhukov et al. (2010) showed that the transformed estimates have the same first-order properties (asymptotic distribution) as the plug-in estimates. The rearrangement method produces a quick fix of the plug-in estimates $(\breve{F}_{co}^{(0)}, \breve{F}_{co}^{(1)})$ and provides promising empirical properties, however it is difficult to use the method in hypothesis testing for evaluating the asymptotic properties.

We want to note that the estimators $\breve{\phi}_{co}, \breve{\phi}_{nt}, \breve{\phi}_{at}$ are driven under the intrinsic assumption that the compliance group membership S_i is independent of Z_i. However, in practice, the sample proportion $\sum_i \mathbf{1}\{S_i = co, Z_i = 0\}/\sum_i \mathbf{1}\{Z_i = 0\}$ might not be the same as $\sum_i \mathbf{1}\{S_i = co, Z_i = 1\}/\sum_i \mathbf{1}\{Z_i = 1\}$ because of sampling variability. In such cases, it would be better to consider compliance class information as additional parameters in order to increase accuracy during estimation and power during hypothesis testing. In the next section, we propose the BL approach incorporating compliance class parameters.

Remark 1 The plug-in estimators $\breve{F}_{co}^{(0)}(t)$ and $\breve{F}_{co}^{(0)}(t)$ are obtained as (3). Other plug-in estimators are $\breve{F}_{at}(t) = \overline{F}_{01}(t)$ and $\breve{F}_{nt}(t) = \overline{F}_{10}(t)$. We denote the vector of the plug-in estimators of the outcome distribution functions, as $\breve{\boldsymbol{F}}(t) = (\breve{F}_{co}^{(0)}(t), \breve{F}_{nt}(t), \breve{F}_{co}^{(1)}(t), \breve{F}_{at}(t))$. We consider the plug-in estimators of the compliance classes as a function of t such that $\breve{\boldsymbol{\phi}}(t) = (\breve{\phi}_{nt}(t), \breve{\phi}_{at}(t))$ where $\breve{\phi}_{nt}(t) = \breve{\phi}_{nt} = n_{10}/(n_{10} + n_{11})$ and $\breve{\phi}_{at}(t) = \breve{\phi}_{at} = n_{01}/(n_{00} + n_{01})$ for all t with terminology slightly abused.

3 The binomial likelihood (BL) approach

3.1 Constructing binomial likelihood with an instrumental variable

Define $\boldsymbol{\theta} : \mathbb{R} \to [0,1]^4$ such that $\boldsymbol{\theta}(t) = (\theta_{co}^{(0)}(t), \theta_{nt}(t), \theta_{co}^{(1)}(t), \theta_{at}(t))$, where $\theta_{co}^{(0)}$, θ_{nt}, $\theta_{co}^{(1)}$, θ_{at}: $\mathbb{R} \to [0,1]$ are functional variables representing four different outcome distributions. Instead of using previously defined $\boldsymbol{F}$, we use the parameter set $\boldsymbol{\theta}$ to emphasize the fact that it is a variable to be estimated. Similarly, we can define the parameter $\boldsymbol{\chi} : \mathbb{R} \to [0,1]^2$ such that $\boldsymbol{\chi}(t) = (\chi_{nt}(t), \chi_{at}(t))$, where $\chi_{nt}(t)$ and $\chi_{at}(t)$ are functional variables representing the proportions of compliance classes. Since the true proportions ϕ_{nt} and ϕ_{at} do not depend on knots, we take the average across the knots to build estimators for them. We set knots $\boldsymbol{t} = (t_1, \ldots, t_m)$ that are the locations to evaluate BL functions later. Then, $\sum_{j=1}^m \chi_{nt}(t_j)/m$ and $\sum_{j=1}^m \chi_{at}(t_j)/m$ are the estimators of ϕ_{nt} and ϕ_{at} respectively. Also, we define $\chi_{co}(t_j) = 1 - \chi_{nt}(t_j) - \chi_{at}(t_j)$, and then $\sum_{j=1}^m (1 - \chi_{nt}(t_j) - \chi_{at}(t_j))/m$ is the estimator of ϕ_{co}. Furthermore, we use χ_1 that is the estimator of ϕ_1. Finally, we define $\theta_{zd}(t_j)$ that is the estimator of $F_{zd}(t_j)$. For example, $\theta_{00}(t_j) = (\chi_{co}(t_j)\theta_{co}^{(0)}(t_j) + \chi_{nt}(t_j)\theta_{nt}(t_j))/(1 - \chi_{at}(t_j))$, $\theta_{01}(t_j) = \theta_{at}(t_j)$, $\theta_{10}(t_j) = \theta_{nt}(t_j)$ and $\theta_{11}(t_j) = (\chi_{co}(t_j)\theta_{co}^{(1)}(t_j) + \chi_{at}(t_j)\theta_{at}(t_j))/(1 - \chi_{nt}(t_j))$.

Denote the data $\mathcal{D}_n = (\boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{Y})$ where $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$, $\boldsymbol{D} = (D_1, \ldots, D_n)^T$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Also, denote the event $K_{zd}^{ij} = \{Z_i = z, D_i = d \mid t_j\}$. The event itself does not depend on the knot t_j, but the probability of occurring this event does depend on t_j through $\boldsymbol{\chi}$. The probability $\mathbb{P}(K_{zd}^{ij})$ can be easily computed in therms of the variables $(\boldsymbol{\chi}, \chi_1)$, discussed in Appendix A.2. At each knot t_j, we define the *point-knot-specific BL function* for a data point (Z_i, D_i, Y_i),

$$L_{ij}(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n) = \prod_{z,d \in \{0,1\}} \mathbb{P}\left(K_{zd}^{ij}\right)^{\mathbf{1}\left(K_{zd}^{ij}\right)} \times \left(\theta_{zd}(t_j)^{\mathbf{1}(Y_i \le t_j)} \left(1 - \theta_{zd}(t_j)\right)^{\mathbf{1}(Y_i > t_j)}\right).$$

Then, by aggregating the point-knot-specific BL functions across all data points, we can define the *knot-specific BL function* at knot t_j,

$$L_j(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n) = \prod_{i=1}^n L_{ij}(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n).$$

Finally, we can define the *BL function* by taking the geometric mean of the knot-specific BL functions across all knots,

$$L(\boldsymbol{\theta}, \chi, \chi_1 | \mathcal{D}_n) = \prod_{j=1}^{m} L_j(\boldsymbol{\theta}, \chi, \chi_1 | \mathcal{D}_n)^{1/m}. \tag{4}$$

The BL function depends on the choice of knots even when the data points are fixed. The knots can be given by researchers, but we propose to use all observed outcomes as knots. More specifically, we use the order statistics $Y_{(j)}$ as knots with $m = n$. This selection procedure provides an automatic way to build the BL function and avoids an arbitrary decision that may cause a favorable conclusion. The contributions of the knot-specific BL functions are, obviously, not independent on data points. Nevertheless, we pretend they are independent. To reduce such dependency, a random sample from $\boldsymbol{Y}$ can be chosen as knots in practice. Also, for a large n, the size of knots does not need to be n. A smaller set of knots can be helpful for reducing computation time. Although we choose $t_j = Y_{(j)}$, we emphasize that, for general knots $\boldsymbol{t}$, the BL function can be constructed and also the MBL estimator can be obtained. Theoretical results in the following section are derived for knots $t_j = Y_{(j)}$. As long as the distribution of knots $\boldsymbol{t}$ is the same as the distribution of $\boldsymbol{Y}$, the theoretical arguments hold.

Remark 2 The knot-specific BL function at knot $t_j = Y_{(j)}$ for some j becomes zero when any of $\theta_{zd}(Y_{(j)})$ is either 0 or 1. This occasionally occurs at the extreme order statistics. To avoid technicalities in the proofs arising from this, we define the likelihood function (4) over the central order statistics, that is, for $j \in I_\kappa = [\lceil n\kappa \rceil, \lceil n(1-\kappa) \rceil]$ for a small fixed constant κ. Throughout the proofs in the supplementary materials, the asymptotics will be in the regime where the sample size n grows to infinity, keeping κ fixed. We omit dependence on κ in the BL for notational brevity. Also, in practice, to avoid computational issues, we can let the knot-specific BL values be 1 when probabilities vanish on the boundary.

3.2 The maximum binomial likelihood (MBL) method

In Sect. 3.1, we introduce a BL approach for constructing a nonparametric likelihood function. Given the BL function $L(\boldsymbol{\theta}, \chi, \chi_1 | \mathcal{D}_n)$, we propose the maximum binomial likelihood (MBL) method to obtain the estimates of $(\boldsymbol{\theta}, \chi, \chi_1)$ by maximizing them over their parameters spaces. To this end, denote $\mathcal{P}([0,1]^{\mathbb{R}})$ as the space of all distribution functions from $\mathbb{R} \to [0,1]$. Let $\boldsymbol{\vartheta}_+ = \{(\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at}) : \theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at} \in \mathcal{P}([0,1]^{\mathbb{R}})\}$, and $\boldsymbol{\varphi}_+ = \{(\chi_{nt}, \chi_{at}) :$ for any $t, (\chi_{nt}(t), \chi_{at}(t)) \in [0,1]^2, 0 \le \chi_{nt}(t) + \chi_{at}(t) \le 1\}$ be the parameter spaces for $\boldsymbol{\theta}$ and χ.

Definition 1 The *MBL* estimate $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}, \hat{\phi}_1)$ is defined as

$$\left(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}, \hat{\phi}_1\right) = \arg \max_{\boldsymbol{\theta}\in\boldsymbol{\vartheta}_+, \boldsymbol{\chi}\in\boldsymbol{\varphi}_+, \chi_1\in[0,1]} L\big(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n\big), \tag{5}$$

where $\hat{\boldsymbol{F}} = (\hat{F}^{(0)}_{co}, \hat{F}_{nt}, \hat{F}^{(1)}_{co}, \hat{F}_{at})$ and $\hat{\boldsymbol{\phi}} = (\hat{\phi}_{nt}, \hat{\phi}_{at})$ are defined at the knots $\boldsymbol{t} = (t_1, \ldots, t_m)$.

Remark 3 The complete parameter space $\boldsymbol{\vartheta}_+ \times \boldsymbol{\varphi}_+ \times [0, 1]$ of the three parameters $(\boldsymbol{F}, \boldsymbol{\phi}, \phi_1)$ will be hereafter referred to as the *restricted parameter space*. To ensure that (5) is well-defined, we extend $\hat{\boldsymbol{F}}$ between the knots by using coordinate-wise right-continuous interpolation and extrapolation beyond the knots by 0 or 1. Also, $(1/m)\sum_{j=1}^m \hat{\phi}_{nt}(t_j)$ and $(1/m)\sum_{j=1}^m \hat{\phi}_{at}(t_j)$ are the estimators of ϕ_{nt} and ϕ_{at}.

The full expression of the binomial log-likelihood function $\ell(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n) = \log L(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n)$ is long and unwieldy. However, we can rewrite it in a compact and instructive form, by grouping and rearranging the terms. It follows (see proof of Proposition 1 below for details) that $\ell(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 | \mathcal{D}_n) = \ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \boldsymbol{\chi}) + \ell_{\boldsymbol{Z}}(\chi_1)$, where $\ell_{\boldsymbol{Z}}(\chi_1) = \frac{1}{n}\big\{(n_{00} + n_{01})\log(1 - \chi_1) + (n_{10} + n_{11})\log\chi_1\big\}$, and

$$\ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \boldsymbol{\chi}) = \sum_{z,d\in\{0,1\}} \ell_{zd}(\boldsymbol{\theta}, \boldsymbol{\chi}),$$

with $\ell_{zd}(\boldsymbol{\theta}, \boldsymbol{\chi})$, for $z, d \in \{0, 1\}$, defined as follows:

$$\ell_{00}(\boldsymbol{\theta}, \boldsymbol{\chi}) = \frac{1}{m}\sum_{j=1}^m n_{00}\Big\{\log\big(1 - \chi_{at}(t_j)\big) + J(\overline{F}_{00}(t_j), \theta_{00}(t_j)\Big\},$$

$$\ell_{10}(\boldsymbol{\theta}, \boldsymbol{\chi}) = \frac{1}{m}\sum_{j=1}^m n_{10}\Big\{\log\chi_{nt}(t_j) + J\Big(\overline{F}_{10}(t_j), \theta_{10}(t_j)\Big)\Big\},$$

$$\ell_{01}(\boldsymbol{\theta}, \boldsymbol{\chi}) = \frac{1}{m}\sum_{j=1}^m n_{01}\Big\{\log\chi_{at}(t_j) + J(\overline{F}_{01}(t_j), \theta_{01}(t_j))\Big\},$$

$$\ell_{11}(\boldsymbol{\theta}, \boldsymbol{\chi}) = \frac{1}{m}\sum_{j=1}^m n_{11}\Big\{\log\big(1 - \chi_{nt}(t_j)\big) + J\Big(\overline{F}_{11}(t_j), \theta_{11}(t_j)\Big\},$$

where the function $J(x, y) = x\log y + (1 - x)\log(1 - y)$.

Proposition 1 *Let $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}, \hat{\phi}_1)$ be the binomial likelihood estimates as defined in (5). Then $\hat{\phi}_1 = \frac{n_{10}+n_{11}}{n}$ that is equal to the plug-in estimate $\breve{\phi}_1$, and*

$$\left(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}\right) = \arg \max_{\boldsymbol{\theta}\in\boldsymbol{\vartheta}_+, \boldsymbol{\chi}\in\boldsymbol{\varphi}_+} \ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \boldsymbol{\chi}).$$

Proof See Section A in the Supplementary Material. □

This proposition shows that the MBL estimate of ϕ_1 is the proportion of individuals with instrument (that is, $Z = 1$) in the observed sample. Furthermore, the MBL estimates of $\boldsymbol{F}$ and $\boldsymbol{\phi}$ can be obtained by maximizing the function $\ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \boldsymbol{\chi})$.

Remark 4 Maximizing the BL function over the *unrestricted parameter space* $\boldsymbol{\vartheta} \times \boldsymbol{\varphi}$, where $\boldsymbol{\vartheta} = \{(\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at}) : \theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at} \in \mathbb{R}^{\mathbb{R}}\}$ with $\mathbb{R}^{\mathbb{R}}$ the set of all functions from $\mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\varphi} = \{(\chi_{nt}, \chi_{at}) :$ for any $t, (\chi_{nt}(t), \chi_{at}(t)) \in \mathbb{R}^2\}$, produces the plug-in estimates $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\vartheta},\boldsymbol{\chi}\in\boldsymbol{\varphi}} \ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \boldsymbol{\chi})$ (see Lemma 1 in the Supplementary Material for the proof).

3.3 Asymptotic properties of the MBL estimates

In this section we discuss the asymptotic properties of the MBL estimates $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$, and how they compare with the plug-in estimates $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$. Assume the knots $t_j = Y_{(j)}$, for $1 \le j \le n$.

Assumption 2 We assume the following:

(a) The proportion parameter vector $\boldsymbol{\phi}$ belongs to the interior of the parameter space $[0, 1]_+^2$.
(b) The distribution functions F_{zd} are continuous, strictly increasing, and have the same support.
(c) For all $K \subset \mathbb{R}$ compact, $s, t \in K$, there exists constants $0 < C_1 \le C_2 < \infty$ (depending on K) such that $C_1|s - t| \le |F_{zd}(s) - F_{zd}(t)| \le C_2|s - t|$.

In particular, Assumption 2 holds whenever F_{zd} are differentiable and the derivatives are uniformly bounded above and below, that is, $C_1 \le F'_{zd}(t) \le C_2$, for all $t \in K$, and $K \subset \mathbb{R}$ compact. Under this assumption we show that the MBL estimates and the plug-in estimates have mean squared errors converging to zero, after rescaling by $\sqrt{n}$. Recall that $H(t) = \sum_{z,d\in\{0,1\}} \eta_{zd} F_{zd}(t)$ is the true population outcome distribution of Y.

Theorem 1 *For any fixed* $0 < \kappa < 1/2$, *let* $I_\kappa = [\lceil n\kappa \rceil, \lceil n(1-\kappa) \rceil]$ *and* $J_\kappa = [H^{-1}(\kappa), H^{-1}(1-\kappa)]$. *Then, the MBL estimates* $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$ *and the plug-in estimates* $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ *satisfy*

$$\frac{1}{|I_\kappa|} \sum_{j\in I_\kappa} \left\| \sqrt{n} \left\{ \hat{\boldsymbol{F}}\big(Y_{(j)}\big) - \breve{\boldsymbol{F}}\big(Y_{(j)}\big) \right\} \right\|_2^2 = o_P(1),$$

and

$$\int_{J_\kappa} \left\| \sqrt{n} \left\{ \hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t) \right\} \right\|_2^2 \mathrm{d}H = o_P(1), \tag{6}$$

where the $o_P(1)$ *term goes to zero as* $n \to \infty$. *Moreover,* $\frac{1}{|I_\kappa|} \sum_{j \in I_\kappa} \| \sqrt{n} \{ \hat{\boldsymbol{\phi}}(Y_{(j)}) - \breve{\boldsymbol{\phi}}(Y_{(j)}) \} \|_2^2 = o_P(1)$. *Also, it implies that the two estimators* $\frac{1}{|I_\kappa|} \sum_{j \in I_\kappa} \hat{\boldsymbol{\phi}}_s(Y_{(j)})$ *and* $\frac{1}{|I_\kappa|} \sum_{j \in I_\kappa} \breve{\boldsymbol{\phi}}_s(Y_{(j)})$ *of the population* ϕ_s *for* $s \in \{nt, at\}$ *satisfy*

$$\sqrt{n} \left(\frac{1}{|I_\kappa|} \sum_{j \in I_\kappa} \hat{\boldsymbol{\phi}}_s(Y_{(j)}) - \frac{1}{|I_\kappa|} \sum_{j \in I_\kappa} \breve{\boldsymbol{\phi}}_s(Y_{(j)}) \right) = o_P(1)$$

Proof See Section B in the Supplementary Material. □

The theorem shows that $\hat{\boldsymbol{F}}$ and $\breve{\boldsymbol{F}}$ (also, $\hat{\boldsymbol{\phi}}$ and $\breve{\boldsymbol{\phi}}$) have the same first-order behavior, and hence the same limiting distribution, which can be derived using the Brownian bridge approximation of the empirical distribution functions; see Corollary 1 (Section C) in the Supplementary Material. Interestingly, Chernozhukov et al. (2010) showed the monotone rearrangement estimates also have the same first-order behavior as the plug-in estimates, which together with Theorem 1, implies that the MBL estimates have the same first-order properties as the rearrangement estimates.

Remark 5 The proof of Theorem 1 can be easily modified to show finite dimensional convergence, that is, for every $s \geq 1$ and given $t_1 < t_2 < \cdots < t_s$, $\| \sqrt{n}(\hat{\boldsymbol{F}}(t_j) - \breve{\boldsymbol{F}}(t_j)) \|_2^2 = o_P(1)$. This would imply that the finite dimensional distributions of the plug-in estimate process $\sqrt{n}(\breve{\boldsymbol{F}}(t) - \boldsymbol{F}(t))$ and the MBL estimate process $\sqrt{n}(\hat{\boldsymbol{F}}(t) - \boldsymbol{F}(t))$ are asymptotically the same. We present this result in terms of mean squared errors as in (6), because it emerges naturally from the asymptotic properties of the BL function, and can be directly applied to the analysis of the BLRT that is introduced in Sect. 4.

4 Extension of the BL approach: hypothesis testing

4.1 Binomial likelihood ratio test (BLRT): Full Version

The BL approach can be extended to constructing a likelihood ratio-type test in a similar way that the ML approach can be extended to constructing a likelihood ratio test. We take two times the difference in two binomial log-likelihood values; one is obtained with the constraint $F_{co}^{(0)}(t) = F_{co}^{(1)}(t)$ (that is, under the null) and the other is obtained without this constraint (that is, under the alternative). This gives a new test for the null hypothesis $H_0 : F_{co}^{(0)}(t) = F_{co}^{(1)}(t)$, and hereafter, we call it the *binomial likelihood ratio test (BLRT)*.

Define the *restricted null parameter space* as $\vartheta_{+,0} = \{(\theta_{co}, \theta_{nt}, \theta_{at}) : \theta_{co}, \theta_{nt}, \theta_{at} \in \mathcal{P}([0,1]^{\mathbb{R}})\}$, where $\mathcal{P}([0,1]^{\mathbb{R}})$ is the set of distribution functions from $\mathbb{R} \to [0,1]$. Then, the BLRT statistic is obtained by

$$T_n = 2\left(\max_{\theta \in \vartheta_+, \chi \in \varphi_+} \ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \chi) - \max_{\theta \in \vartheta_{+,0}, \chi \in \varphi_+} \ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \chi) \right). \tag{7}$$

Let $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}}) = \arg\max_{\theta \in \vartheta_{+,0}, \chi \in \varphi_+} \ell_{\boldsymbol{Y},\boldsymbol{D}|\boldsymbol{Z}}(\boldsymbol{\theta}, \chi)$. The asymptotic properties of $\hat{\boldsymbol{\psi}}$ can be derived as we did for $\hat{\boldsymbol{F}}$ in Theorem 1. Also, we can derive the asymptotically equivalent plugin-type estimators that have not been studied before. It is worth noting that the explicit form of the equivalent estimators of $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ is provided in Section D, the Supplementary Material.

Theorem 2 *Fix* $0 < \kappa < 1/2$ *and recall that* $H(t) = P(Y \le t)$. *Let* T_n *be the binomial likelihood ratio test statistic as defined in* (7). *Denote* $\bar{J}_\kappa = [\overline{H}^{-1}(\kappa), \overline{H}^{-1}(1-\kappa)]$. *Then*,

$$T_n = \frac{n_0 n_1}{n} \int_{\bar{J}_\kappa} \frac{\left(\bar{F}_0(t) - \bar{F}_1(t)\right)^2}{\bar{H}(t)\left(1 - \bar{H}(t)\right)} \mathrm{d}\bar{H}(t) + o_P(1),$$

where $\bar{H}(t) = (n_0 \bar{F}_0(t) + n_1 \bar{F}_1(t))/n$ *is the empirical distribution function of* Y.

Proof See Section E in the Supplementary Material. □

This theorem gives an asymptotically equivalent representation of the BLRT statistic as the two-sample Anderson–Darling test statistic (Pettitt, 1976). It can be used to construct the rejection region and compute the critical value for a given significance level. Moreover, this shows that the test based on T_n is consistent against all fixed alternatives, because of the universal consistency of the two-sample Anderson–Darling test (Scholz & Stephens, 1987).

However, in finite-sample settings, the critical value obtained from the asymptotic distribution of T_n can be conservative. In the theorem above, to derive the asymptotic properties, we use the equivalent plug-in estimators of $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ instead of using $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\xi}})$ directly. However, they do not lie in the restricted parameter space, which leads to a gap between the equivalent and actual BL values. This gap fades out as n increases, but it can be critical when we evaluate finite-sample performance. This issue will be further discussed in the simulation section.

The BLRT is developed for testing the null hypothesis $H_0 : F_{co}^{(0)}(t) = F_{co}^{(1)}(t)$ that assumes no treatment effect for compliers. The BLRT can be further extended to testing other hypotheses like $H_0^g : F_{co}^{(0)}(g(t)) = F_{co}^{(1)}(t)$ for some g. To test H_0^g, a simple modification is required. A new outcome variable Y_i^* can be generated: $Y_i^* = g(Y_i)$ if $D_i = 1$, and $Y_i^* = Y_i$ otherwise. Then, (Z_i, D_i, Y_i^*) can be used for the BLRT as (7), and this test based on the new dataset conducts a hypothesis test for H_0^g. Among many choices of g, $g(t) = t - \mu$ can be considered to check whether there is any location shift between the two distributions. The assumption of the location shift means

that there is a constant treatment effect μ for all compliers. Therefore, this test can be used for examining treatment effect heterogeneity. If $H_0^{location} : F_{co}^{(0)}(t-\mu) = F_{co}^{(1)}(t)$ is rejected for all μ, then there is evidence that treatment effects are heterogeneous.

4.2 Binomial likelihood ratio test: simple version

As we discussed in Sect. 2.2, under Assumption 1, testing the null hypothesis $H_0 : F_{co}^{(0)}(t) = F_{co}^{(1)}(t)$ is equivalent to testing the null hypothesis $H_0^{simple} : F_0(t) = F_1(t)$. Based on this, we propose a simple version of the BLRT by comparing $F_0(t) = F_1(t)$ instead of $F_{co}^{(0)}(t) = F_{co}^{(1)}(t)$. This test does not use the information of compliance classes by ignoring the treatment $\boldsymbol{D}$, but uses only $\boldsymbol{Z}$ and $\boldsymbol{Y}$. We define the parameter $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}(t) = (\theta_0(t), \theta_1(t))$, where $\theta_0(t), \theta_1(t) : \mathbb{R} \to [0,1]$ are functional variables representing the outcome distributions for $Y|Z=0$ and $Y|Z=1$. Then, given that $\boldsymbol{Z}$ is conditioned on, the *simple version binomial log-likelihood function* $\ell_{\boldsymbol{Y}|\boldsymbol{Z}}^{simple}(\boldsymbol{\theta})$ is

$$\begin{aligned}\ell_{\boldsymbol{Y}|\boldsymbol{Z}}^{simple}(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{1}\big(Z_i = 0, Y_i \le t_j\big)\log\theta_0\big(t_j\big) \\ &\quad + \mathbf{1}\big(Z_i = 0, Y_i \le t_j\big)\log\big(1-\theta_0\big(t_j\big)\big) \\ &\quad + \mathbf{1}\big(Z_i = 1, Y_i \le t_j\big)\log\theta_1\big(t_j\big) \\ &\quad + \mathbf{1}\big(Z_i = 1, Y_i \le t_j\big)\log\big(1-\theta_1\big(t_j\big)\big) \\ &= \frac{1}{m}\sum_{j=1}^{m} n_0 J\Big(\overline{F}_0\big(t_j\big), \theta_0(t_j)\Big) + n_1 J\Big(\overline{F}_1\big(t_j\big), \theta_1\big(t_j\big)\Big),\end{aligned} \tag{8}$$

where $J(x,y) = x\log y + (1-x)\log(1-y)$. The *simple version BLRT statistic* T_n^{simple} is defined as

$$T_n^{simple} = 2\left(\max_{\theta_0,\theta_1\in\mathcal{P}([0,1]^{\mathbb{R}})} \ell_{\boldsymbol{Y}|\boldsymbol{Z}}^{simple}(\boldsymbol{\theta}) - \max_{\theta_0=\theta_1\in\mathcal{P}([0,1]^{\mathbb{R}})} \ell_{\boldsymbol{Y}|\boldsymbol{Z}}^{simple}(\boldsymbol{\theta})\right).$$

Since this test does not use any information of the compliance class behaviors, it does not require estimation of the proportions, and estimation of the outcome distribution for the compliance classes. The following gives the asymptotic approximation of T_n^{simple}.

Theorem 3 *The test statistic T_n^{simple} has an explicit form as*

$$T_n^{simple} = 2\Big(\ell_{\boldsymbol{Y}|\boldsymbol{Z}}^{simple}\Big(\overline{F}_0, \overline{F}_1\Big) - \ell_{\boldsymbol{Y}|\boldsymbol{Z}}^{simple}\Big(\overline{H}, \overline{H}\Big)\Big).$$

If we assume that the knots are $\boldsymbol{t} = (Y_{(1)}, \ldots, Y_{(n)})$, then the test statistic T_n^{simple} is asymptotically equivalent to the two-sample Anderson–Darling test statistic, that is,

$$T_n^{simple} = \frac{n_0 n_1}{n} \int_{\bar{J}_\kappa} \frac{\left(\bar{F}_0(t) - \bar{F}_1(t)\right)^2}{\bar{H}(t)\left(1 - \bar{H}(t)\right)} \mathrm{d}\bar{H}(t) + o_P(1).$$

Proof See Section E in the Supplementary Material. □

Theorem 3 shows that, as in the case of the full version BLRT, the simple version BLRT is asymptotically equivalent to the two-sample Anderson–Darling test, and, therefore, is consistent against all fixed alternatives, as well. The difference is that T_n^{simple} has a closed form and does not need the EM-PAV algorithm that will be introduced in the next section. However, T_n^{simple} does not involve any estimation procedure of outcome distributions for compliance classes, and, hence, cannot be applied for estimation purposes.

5 Computation and simulation

5.1 EM-PAV algorithm for computing the MBL estimates

There are no closed form solutions to the MBL estimates. However, the estimates can be computed efficiently by using a combination of the expectation-maximization (EM) algorithm and the pool-adjacent-violator(PAV) algorithm. We call it the *EM-PAV* algorithm. To begin with, we introduce the *complete-data* $\overline{\mathcal{D}}_n$, which includes the compliance class $\boldsymbol{S}$, $\overline{\mathcal{D}}_n = (\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{D}, \boldsymbol{Y})^T$. If Z_i and S_i are known, then D_i is determined; for example, if $Z_i = 0$ and $S_i = co$, then $D_i = 0$. Denote the event $\overline{K}_{zs}^{ij} = \{Z_i = z, S_i = s \mid t_j\}$, where $s \in \{co, at, nt\}$.

Given the complete data, we can define a *point-knot-specific complete-data binomial likelihood function* for the data point (Z_i, S_i, D_i, Y_i) at knot t_j,

$$\overline{L}_{ij}\left(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 \mid \overline{\mathcal{D}}_n\right) = \prod_{z\in\{0,1\}} \prod_{s\in\{co,nt,at\}} \mathbb{P}\left(\overline{K}_{zs}^{ij}\right)^{\mathbf{1}\left(\overline{K}_{zs}^{ij}\right)} \times \left(\theta_s(t_j)^{\mathbf{1}(Y_i \le t_j)} \left(1 - \theta_s(t_j)\right)^{\mathbf{1}(Y_i > t_j)}\right).$$

The *complete-data binomial likelihood* is obtained by combining all point-knot-specific complete-data likelihood functions in the same way to define the BL,

$$\overline{L}\left(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 \mid \overline{\mathcal{D}}_n\right) = \prod_{j=1}^{m} \left\{ \prod_{i=1}^{n} \overline{L}_{ij}\left(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 \mid \overline{\mathcal{D}}_n\right) \right\}^{1/m}.$$

As in the BL, the dependence on χ_1 in the complete-data likelihood is separable, that is,

$$\log \overline{L}\left(\boldsymbol{\theta}, \boldsymbol{\chi}, \chi_1 \mid \overline{\mathcal{D}}_n\right) = \ell(\chi_1) + \log \overline{L}\left(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \overline{\mathcal{D}}_n\right)$$

where $\ell(\chi_1) = (n_{00} + n_{01}) \log(1 - \chi_1) + (n_{10} + n_{11}) \log \chi_1$, and $\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \overline{\mathcal{D}}_n)$ does not depend on χ_1. Hereafter, we will refer to $\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \overline{\mathcal{D}}_n)$ as the *complete-data binomial log-likelihood*. To find the maximizer of $\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \overline{\mathcal{D}}_n)$, our algorithm is initiated by specifying the initial values $(\boldsymbol{\theta}_{(0)}, \boldsymbol{\chi}_{(0)})$ that lie in the parameter space $\boldsymbol{\vartheta}_+ \times \boldsymbol{\varphi}_+$. Then the following steps are repeated until the values converge:

Algorithm 1 (*EM-PAV algorithm*) Let $\hat{\boldsymbol{\theta}}_{(k)} = (\hat{\boldsymbol{\theta}}^{(0)}_{co,(k)}, \hat{\boldsymbol{\theta}}_{nt,(k)}, \hat{\boldsymbol{\theta}}^{(1)}_{co,(k)}, \hat{\boldsymbol{\theta}}_{at,(k)})$ and $\hat{\boldsymbol{\chi}}_{(k)} = (\hat{\boldsymbol{\chi}}_{nt,(k)}, \hat{\boldsymbol{\chi}}_{at,(k)})$ be the outputs after the kth step of the iteration. The following shows the $(k + 1)$th step.

(*Expectation Step*) Given these outputs $(\hat{\boldsymbol{\theta}}_{(k)}, \hat{\boldsymbol{\chi}}_{(k)})$and the observed data $\mathcal{D}_n$, the expected complete-data binomial log-likelihood is

$$Q_k(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \hat{\boldsymbol{\theta}}_{(k)}, \hat{\boldsymbol{\chi}}_{(k)}) = E_{\hat{\boldsymbol{\theta}}_{(k)}, \hat{\boldsymbol{\chi}}_{(k)}} \Big(\log \overline{L}\Big(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \overline{\mathcal{D}}_n \Big) \mid \mathcal{D}_n \Big).$$

The expectation can be easily calculated; see Section F2 in the Supplementary Material for computational details.

(*Maximization Step*) To begin with, define

$$\big(\breve{\boldsymbol{\theta}}_{(k+1)}, \breve{\boldsymbol{\chi}}_{(k+1)}\big) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}, \boldsymbol{\chi} \in \boldsymbol{\varphi}} Q_k\big(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \hat{\boldsymbol{\theta}}_{(k)}, \hat{\boldsymbol{\chi}}_{(k)}\big).$$

Note that $\breve{\boldsymbol{\theta}}_{(k+1)} = \Big(\breve{\boldsymbol{\theta}}^{(0)}_{co,(k+1)} \breve{\boldsymbol{\theta}}_{nt,(k+1)}, \breve{\boldsymbol{\theta}}^{(1)}_{co,(k+1)}, \breve{\boldsymbol{\theta}}_{at,(k+1)}\Big)$, where $\breve{\boldsymbol{\theta}}^{(0)}_{co,(k+1)}$ is evaluated at knots $Y_{(j)}$, and similarly for other estimates. Observe that $(\breve{\boldsymbol{\theta}}_{(k+1)}, \breve{\boldsymbol{\chi}}_{(k+1)})$ is the unrestricted maximizer of $Q_k(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \hat{\boldsymbol{\theta}}_{(k)}, \hat{\boldsymbol{\chi}}_{(k)})$. These estimates can be computed explicitly; see Section F3 in the Supplementary Material. It can be shown that $\breve{\boldsymbol{\chi}}_{(k+1)} = (\breve{\boldsymbol{\chi}}_{nt,(k+1)}, \breve{\boldsymbol{\chi}}_{at,(k+1)})$ is actually in the restricted space $\boldsymbol{\varphi}_+$, that is, $\breve{\boldsymbol{\chi}}_{nt,(k+1)}(t_j), \breve{\boldsymbol{\chi}}_{at,(k+1)}(t_j) \in [0, 1]$ and $0 \le \breve{\boldsymbol{\chi}}_{nt,(k+1)}(t_j) + \breve{\boldsymbol{\chi}}_{at,(k+1)}(t_j) \le 1$ for any knot t_j. Define $\hat{\boldsymbol{\chi}}_{(k+1)} = \breve{\boldsymbol{\chi}}_{(k+1)}$. In general, however, $\breve{\boldsymbol{\theta}}_{(k+1)} \notin \boldsymbol{\vartheta}_+$, because $\breve{\boldsymbol{\theta}}_{(k+1)}$ may not satisfy the non-decreasing condition of distribution functions. To ensure the monotonicity constraint we apply the PAV algorithm to the estimate $\breve{\boldsymbol{\theta}}_{(k+1)}$,

$$\hat{\boldsymbol{\theta}}_{(k+1)} = \mathrm{PAV}_{\boldsymbol{w}}(\breve{\boldsymbol{\theta}}^{(0)}_{co,(k+1)}, \breve{\boldsymbol{\theta}}_{nt,(k+1)}, \breve{\boldsymbol{\theta}}^{(1)}_{co,(k+1)}, \breve{\boldsymbol{\theta}}_{at,(k+1)}),$$

where the operation $\mathrm{PAV}_{\boldsymbol{w}}$ is applied coordinate-wise and the weight vector is $\boldsymbol{w}_{(k+1)} = (\boldsymbol{w}^{(0)}_{co,(k+1)}, \boldsymbol{w}_{nt,(k+1)}, \boldsymbol{w}^{(1)}_{co,(k+1)}, \boldsymbol{w}_{at,(k+1)})$, where the weights are defined in Section F in the Supplementary Material.

The following proposition establishes the correctness of this algorithm.

Proposition 2 *Let $\hat{\boldsymbol{\theta}}_{(k+1)}, \hat{\boldsymbol{\chi}}_{(k+1)}$ be as defined above. Then*

$$\big(\hat{\boldsymbol{\theta}}_{(k+1)}, \hat{\boldsymbol{\chi}}_{(k+1)}\big) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+, \boldsymbol{\chi} \in \boldsymbol{\varphi}_+} Q_k\big(\boldsymbol{\theta}, \boldsymbol{\chi} \mid \hat{\boldsymbol{\theta}}_{(k)}, \hat{\boldsymbol{\chi}}_{(k)}\big).$$

Proof See Section F in the Supplementary Material. □

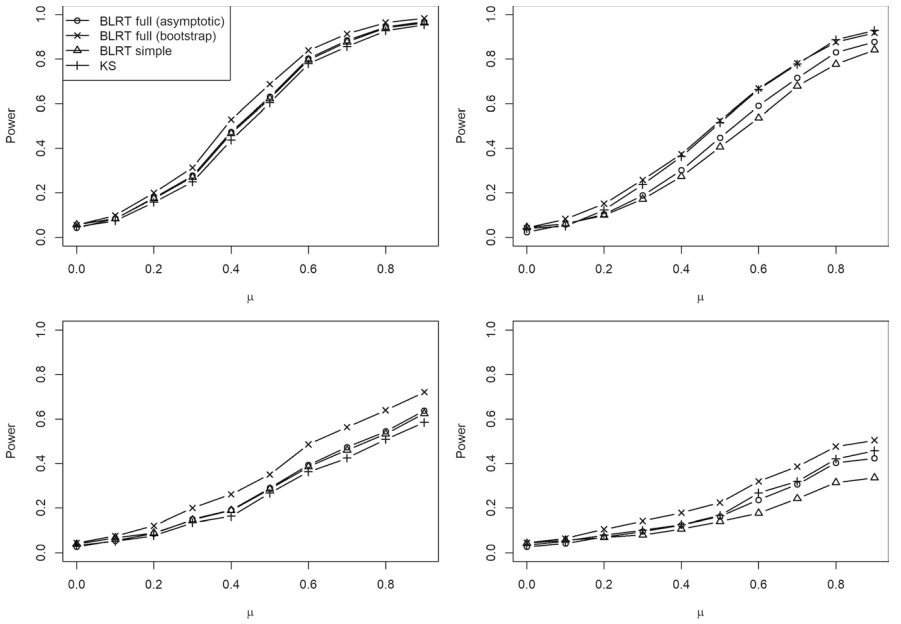

Fig. 1 Power of T_n, T_n^{simple} and T_{KS}. Power is calculated given a significance level $\alpha = 0.05$. Upper left: $(\mu_{nt}, \mu_{at}) = (-1, 1)$ and strong IV, Upper right: $(\mu_{nt}, \mu_{at}) = (-2, 2)$ and strong IV, Lower left: $(\mu_{nt}, \mu_{at}) = (-1, 1)$ and weak IV, lower right: $(\mu_{nt}, \mu_{at}) = (-2, 2)$ and weak IV

Table 1 Size and power of the different tests with a significance level 0.05

(μ_{nt}, μ_{at})	IV	μ	$N(-\mu, 1)$ vs. $N(\mu, 1)$				
			T_n (boot.)	T_n (asymp.)	T_n^{simple}	T_{AD}	T_{KS}
$(-1, 1)$	Strong	0	0.054	0.042	0.056	0.047	0.047
		0.3	0.313	0.277	0.270	0.287	0.249
		0.6	0.839	0.802	0.796	0.777	0.779
		0.9	0.983	0.968	0.963	0.971	0.954
	Weak	0	0.044	0.027	0.053	0.052	0.050
		0.3	0.200	0.149	0.146	0.120	0.134
		0.6	0.486	0.393	0.385	0.356	0.362
		0.9	0.721	0.637	0.624	0.646	0.584
$(-2, 2)$	Strong	0	0.041	0.022	0.047	0.050	0.046
		0.3	0.258	0.188	0.171	0.170	0.237
		0.6	0.668	0.590	0.536	0.529	0.663
		0.9	0.918	0.877	0.841	0.831	0.928
	Weak	0	0.045	0.026	0.042	0.043	0.033
		0.3	0.142	0.095	0.080	0.081	0.101
		0.6	0.319	0.236	0.177	0.185	0.267
		0.9	0.504	0.423	0.336	0.339	0.457

5.2 Simulation: performance of BLRT

To assess the performance of the two versions of the proposed BLRT, we compare them to the Kolmogorov–Smirnov test with T_{KS} in a simulation study. Note that both T_n^{simple} and T_{KS} do not use the variable $\boldsymbol{D}$, but use $\boldsymbol{Z}$ and $\boldsymbol{Y}$. The null distribution of T_{KS} can be obtained by permuting $\boldsymbol{Z}$ multiple times while the Anderson–Darling distribution A^2 is used as the asymptotic null distribution of T_n^{simple}. The distribution A^2 is the limiting distribution of the two-sample Anderson–Darling test statistic T_{AD} (Pettitt, 1976). Also, A^2 is used as the limiting distribution of the full version BLRT T_n as well.

In the simulation study, assume that all four potential outcome distributions are normal distributions with variance 1, but with different means: $F_{co}^{(0)} \sim N(\mu_{co}^{(0)}, 1)$, $F_{co}^{(1)} \sim N(\mu_{co}^{(1)}, 1)$, $F_{nt} \sim N(\mu_{nt}, 1)$ and $F_{at} \sim N(\mu_{at}, 1)$. Two simulation factors are considered: (1) how far the distributions are from each other, (2) how strong the IV is. To see the impact of the first factor, we consider two simulation settings with $(\mu_{nt}, \mu_{at}) = (-1, 1)$ (close) and $(\mu_{nt}, \mu_{at}) = (-2, 2)$ (far). We evaluate these settings with $(\mu_{co}^{(0)}, \mu_{co}^{(1)}) = (-\mu, \mu)$ for various μ values. In addition, to assess the second factor, we consider the weak IV setting with the proportions $(\phi_{co}, \phi_{nt}, \phi_{at}) = (0.2, 0.4, 0.4)$ and the strong IV setting with $(\phi_{co}, \phi_{nt}, \phi_{at}) = (1/3, 1/3, 1/3)$. We consider four simulations settings, and, in each simulation setting, various values of μ and n are considered.

Table 1 shows estimated size and power of T_n, T_n^{simple}, T_{AD} and T_{KS} from 1000 simulated datasets. This table reports the four simulation settings for $n = 300$ and $\mu = (0, 0.3, 0.6, 0.9)$. Other values of μ are not reported in this table, but are plotted in Fig. 1. More simulation results are reported in the Supplementary Materials (Section G). The first row of each simulation setting shows the simulated size. For power comparisons, one of the main findings is that the shape difference between $F_0(t)$ and $F_1(t)$ is important for the performances of the tests. When $(\mu_{nt}, \mu_{at}) = (-1, 1)$, $F_0(t)$ and $F_1(t)$ differ at tails, and T_n and T_n^{simple} outperforms T_{KS}. However, when $(\mu_{nt}, \mu_{at}) = (-2, 2)$, $F_0(t)$ and $F_1(t)$ differ mostly at the middle, and T_{KS} outperforms the others. For another finding, when an IV is weak, meaning that ϕ_{co} is small, power is reduced, but at the same time, the shape difference between $F_0(t)$ and $F_1(t)$ is less centered since $F_{nt}(t)$ and $F_{at}(t)$ dominate the shape. Therefore, as ϕ_{co} decreases, T_n can capture the distributional difference more and produce better performance than T_{KS}. This can be found in Fig. 1 by comparing the upper right plot and lower right plot; the asymptotic T_n is less powerful than T_{KS} when an IV is strong, but they have the almost same power in the weak IV setting. In summary, when the difference of the distributions in two samples is not concentrated in the middle, T_n and T_n^{simple} can be powerful.

As we pointed out in the previous section, the simulation results indicate that using the limiting distribution A^2 for T_n is conservative. The simulated sizes do not reach the nominal level 0.05. We conducted additional simulations for various n values when there is no effect at all. We conducted simulations for different sample sizes $n = (500, 1000, 1500, 2000)$ and the estimated sizes from 10,000 simulations

for each n are (0.027, 0.029, 0.032, 0.034) when $(\mu_{nt}, \mu_{at}) = (-2, 2)$. As we expected, the size approaches to the correct nominal level $\alpha = 0.05$ as n increases. However, the convergence for T_n is not satisfactory for a moderately large n. This conservativeness essentially lowers the performance of T_n in finite samples.

To boost the finite-sample performance, we can consider the bootstrapping method that simulates the true null distribution of T_n under the null hypothesis for given n. Bootstrapping can be done using the estimates $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\xi}}$ that are obtained under the null hypothesis. For the bth procedure of bootstrapping, first fix $\boldsymbol{Z}$ and sample the compliance class membership $\boldsymbol{S}^{(b)}$ using $\hat{\boldsymbol{\xi}}$. Second, determine $\boldsymbol{D}^{(b)}$ based on $\boldsymbol{Z}$ and $\boldsymbol{S}^{(b)}$; for instance, if $Z_i = 0$ and $S_i^{(b)} = co$, then $D_i^{(b)} = 0$. Third, take a sample $\boldsymbol{Y}^{(b)}$ based on $\boldsymbol{Z}$ and $\boldsymbol{S}^{(b)}$ using the estimate $\hat{\boldsymbol{\psi}}$. Finally, repeat the entire process for $1 \le b \le B$ to obtain the bootstrapped samples $\{(\boldsymbol{Z}, \boldsymbol{D}^{(b)}, \boldsymbol{Y}^{(b)})\}_{1 \le b \le B}$. Table 1 reports simulated size and power based on $B = 1000$ bootstrapped samples for each simulated dataset. The column of T_n (boot.) in Table 1 shows the estimated size and power from the bootstrap procedure. All the values are improved from the asymptotic-version values. The bootstrap-version T_n can reduce the performance gap in cases where T_{KS} is superior, and in some cases, can overtake T_{KS}.

5.3 Simulation: performance of the MBL method

In this section, we evaluate the MBL estimates by comparing it with the plug-in estimates (3) and the estimates obtained from the rearrangement method proposed by Chernozhukov et al. (2010).

We consider the four situations in simulation studies. In the first three situations, all distributions are Gamma distributions: $F_{co}^{(0)} = F_{co}^{(1)} \sim Gamma(1.2^2, 1), F_{nt} \sim Gamma(1^2, 1)$ and $F_{at} \sim Gamma(1.4^2, 1)$.

Table 2 Average performance of the three estimation methods

Situation	Method	Bias	SE	1000MSE
1	Plug-in	0.0514	0.0923	11.16
	MBL	0.0288	0.0330	1.92
	Rearrangement	0.0359	0.0740	6.76
2	Plug-in	0.0098	0.0101	0.20
	MBL	0.0088	0.0089	0.16
	Rearrangement	0.0087	0.0094	0.16
3	Plug-in	0.0032	0.0032	0.02
	MBL	0.0031	0.0031	0.02
	Rearrangement	0.0030	0.0031	0.02
4	Plug-in	0.0612	0.1745	34.20
	MBL	0.0276	0.0328	1.84
	Rearrangement	0.0255	0.0404	2.28

The compliance class proportions are (1) $(\phi_{co}, \phi_{nt}, \phi_{at}) = (0.10, 0.45, 0.45)$, (2) $(\phi_{co}, \phi_{nt}, \phi_{at}) = (0.2, 0.4, 0.4)$, (3) $(\phi_{co}, \phi_{nt}, \phi_{at}) = (1/3, 1/3, 1/3)$. In the fourth situation, all distributions are normal distributions: $F_{co}^{(0)} = F_{co}^{(1)} \sim N(0, 1)$, $F_{nt} \sim N(-1, 1)$ and $F_{at} \sim N(1, 1)$ with $(\phi_{co}, \phi_{nt}, \phi_{at}) = (0.10, 0.45, 0.45)$. The sample size is $n = 1000$. To compute the average performance, we consider 1000 simulated datasets. For each dataset, we compute the L_2 distance between the estimated function $\hat{F}$ and the true function F, $L_2(\hat{F}, F) = \int (\hat{F} - F)^2 \mathrm{d}F$.

Table 2 shows the average performance of three considered estimation methods. Biases, standard errors and mean squared errors are reported. The MBL method has the least bias in situation 1, but the rearrangement method has the least bias in situations 2, 3 and 4. However, the MBL method has the least standard errors in every situation. Moreover, it has the best mean squared error in all the situations, although it has similar performance to the rearrangement method when an instrument is not weak.

6 Oregon health insurance experiment: the effect of medicaid coverage on mental health

We consider the 2008 Oregon health insurance experiment data that is publicly available from https://www.nber.org/oregon/1.home.html. To investigate the effect of Medicaid on health outcomes, Oregon opened a waiting list for a limited number of spots in its Medicaid program for low-income, uninsured, able-bodied adults between 19-64, which had previously been closed to new enrollment. From the waiting list, people selected by random lottery drawings, won the opportunity for themselves, and any household member, to apply for Oregon Health Program (OHP) Standard. However, not all persons selected by the lottery enrolled in Medicaid, either because they did not apply or because they were deemed ineligible. The lottery process and OHP standard are described in more detail in Finkelstein et al. (2012). This random assignment embedded in the lottery allows us to study the effect of Medicaid coverage in a random encouragement design. An indicator of winning a lottery is the instrumental variable. Also, enrollment to the Medicaid program is the non-randomized treatment variable. Approximately 2 years after the lottery, health outcomes are measured for persons who responded to the follow-up survey. In our example, we use a self-reported mental health outcome by using scores on the Medical Outcome Study 8-Item Short-Form Survey (SF-8). The scores range from 0 to 100, with higher scores indicating better self-reported health-related quality of life. The scale is normalized to yield a mean of 50 and a standard deviation of 10 in the general U.S. population. Details of other health outcomes and data collection have been provided in Baicker et al. (2013).

From the data, we consider a sample of 1,117 African Americans (single-person households) who signed themselves up for the lottery. Among them, 546 people (48.9%) were selected by the lottery drawings. The probability of Medicaid coverage is 0.511 in the lottery winning group and 0.226 in the other group. The plug-in estimates of the proportions for compliance classes are $(\breve{\phi}_{co}, \breve{\phi}_{nt}, \breve{\phi}_{at}) = (0.285, 0.489, 0.226)$. Lottery selection increased the probability

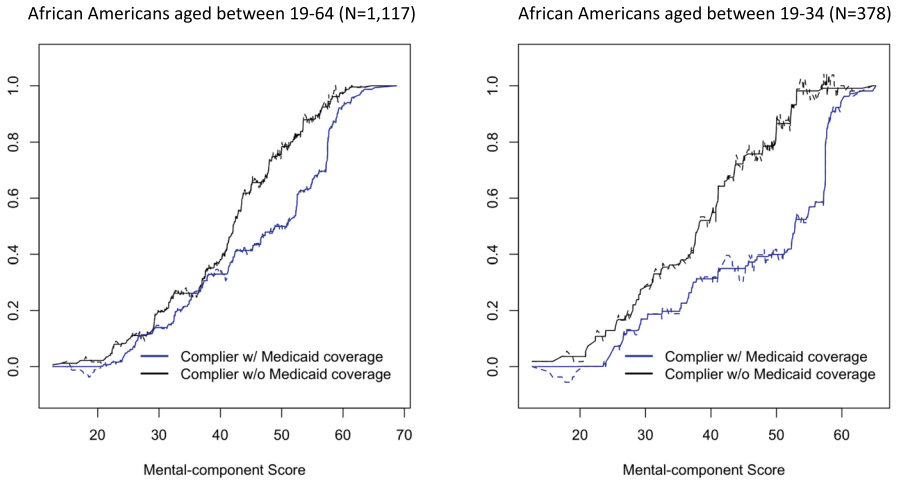

Fig. 2 Estimated distribution functions of the mental-component scores for compliers in the African-American population. Higher mental-component scores indicates better self-reported mental health. The dotted blue and black lines are the plug-in estimates and the solid blue and black lines are the BL estimates of the distribution functions of the complier with Medicaid coverage and the complier without medicaid coverage, respectively

of Medicaid coverage by 28.5 points among the single-person African American households. The MBL estimates of the proportions are the same as the plug-in estimates up to three decimal places. The two-stage least squares (2SLS) estimate is 4.88 (95% CI 0.01– 9.75) with p-value 0.050. The magnitude of improvement was approximately half of the standard deviation of the mental-component score. Furthermore, we can restrict our attention to a subsample of African Americans aged between 19 and 34 ($N = 378$). The 2SLS estimate is 9.92 (95% CI 0.98–18.86) with p-value 0.030. The estimated proportions of compliance classes are (0.284, 0.495, 0.221) which are almost identical to the estimated proportions for the total African American population.

Figure 2 shows the estimated distribution functions of the potential outcomes of mental-component scores for compliers when enrolled in the Medicaid program and when not enrolled. The left plot shows the plug-in estimates described in Sect. 2.2 and the MBL estimates for African Americans aged between 19 and 64 (the full sample), and the right plot shows them for African Americans aged between 19-34. In both of the plots, we see that the estimated distribution function for complier without Medicaid coverage is almost always above the other. The gap between the two functions is wider at higher mental-component scores. Unlike the plug-in estimate, the MBL estimate satisfies the non-decreasing condition, and, as a result, there is a unique value of estimated scores corresponding to a specific quantile level. This feature can be useful for those who want to estimate the treatment effect at a certain quantile level using the estimated distribution functions. For example, the MBL method estimates that Medicaid coverage led to an increase of 8.70 points in the median score on the mental component for compliers. However, from the plug-in method, there are two values that correspond

to the value 0.5 of the distribution for complier with Medicaid coverage, making it unclear how to compare the medians of the two distribution functions. For the young African American population (aged between 19-34), Medicaid coverage increased the median mental-component score by 14.68 points from 38.13 to 52.81.

Furthermore, the BLRT can be conducted for testing the null hypothesis $H_0 : F_{co}^{(0)} = F_{co}^{(1)}$. Our simulation studies suggest that T_n can be more powerful than T_{KS} in such setting. We apply both versions of the proposed binomial likelihood test, T_n and T_n^{simple}, and compare them with T_{KS}. Using the asymptotic null distribution A^2, the P-values are computed as (0.021, 0.020, 0.031) for $(T_n, T_n^{simple}, T_{KS})$. Moreover, the P-value of T_n can be computed by the bootstrap procedure with $B = 10,000$ described in Sect. 5.2, and the P-value is 0.021, which agrees with the asymptotic version of the P-value. Similarly, for the young African American population, the estimated P-values are (0.012, 0.012, 0.030) for $(T_n, T_n^{simple}, T_{KS})$. For a smaller sample size, the proposed BLRT produced a smaller P-value. For all considered tests, we reject equality of distributions at a significance level $\alpha = 0.05$.

Finally, using the BLRT, we also test another hypothesis $H_0^{location} : F_{co}^{(0)}(t - \mu) = F_{co}^{(1)}(t)$ for testing treatment effect heterogeneity. All possible values of μ are examined for the African American population and its subpopulation of African American aged between 19-34. For the African American population, $H_0^{location}$ is not rejected for values between $0.76 \leq \mu \leq 10.59$. Also, for the African American aged between 19-34, $H_0^{location}$ is not rejected for values between $1.92 \leq \mu \leq 21.38$. These results show that there is no evidence of treatment effect heterogeneity.

7 Discussion

We propose a non-parametric composite likelihood approach, referred to as the binomial likelihood (BL) method, for making causal inferences about the distributional treatment effect in a randomized experiment with an instrumental variable. The BL approach provides a non-parametric inferential tool similar to the classical parametric likelihood. The maximum binomial likelihood (MBL) method provides estimates of the outcome distributions, which are proper distribution functions and the binomial likelihood ratio test (BLRT) is a powerful technique to detect distributional changes when the outcome distributions are close to each other, especially when the IV is weak.

Several extensions and generalizations of the BL method are possible. For instance, while constructing the BL functions one requires specification of the knots as evaluating points. We recommended to use all the observed outcomes for the knots, but a different specification can be considered depending on the question of interest. For example, in the Oregon Health Insurance experiment, if one wants to examine whether there is any effect for people above mental score

40, then only outcomes above 40 can be chosen as knots. In such a setting, a rejection implies that there is evidence of distributional changes for this specific subpopulation.

The results in this paper show that the BL approach works well in randomized encouragement experiments where a compliance class is latent. A possible future direction could be to study the performance of the BL approach in general mixture models when there is a latent variable.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42952-023-00233-4.

Funding This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korea government(MSIT) (2021R1C1C1012750).

Data availability statement The data that support the findings of this study is publicly available from https://www.nber.org/oregon/1.home.html.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association, 97*(457), 284–292.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics, 113*(2), 231–263.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444–455.

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., & Finkelstein, A. N. (2013). The Oregon experiment—effects of Medicaid on clinical outcomes. *New England Journal of Medicine, 368*(18), 1713–1722.

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine, 33*(13), 2297–2340.

Brookhart, M. A. & Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics*, *3*(1), 14. https://doi.org/10.2202/1557-4679.1072

Cheng, J., Qin, J., & Zhang, B. (2009). Semiparametric estimation and inference for distributional and general treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*(4), 881–904.

Cheng, J., Small, D. S., Tan, Z., & Ten Have, T. R. (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika, 96*(1), 19–36.

Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica, 78*(3), 1093–1125.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., Oregon Health Study Group. (2012). The Oregon health insurance experiment: evidence from the first year. *The Quarterly Journal of Economics, 127*(3), 1057–1106.

Geman, S., & Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics, 10*(2), 401–414.

Heagerty, P. J., & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association, 93*(443), 1099–1111.

Hernan, M. A., & Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology, 17*(4), 360–372.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology, 18*, 449–484.

Johnson, M., Cao, J., & Kang, H. (2019). Detecting heterogeneous treatment effect with instrumental variables. *arXiv preprint* arXiv:1908.03652.

Kang, H., Peck, L., & Keele, L. (2018). Inference for instrumental variables: A randomization inference approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181*(4), 1231–1254.

Larribe, F., & Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica, 21*(1), 43–69.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics, 80*(1), 221–39.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Roczniki Nauk Roiniczych, X*(5), 1–51. Reprinted in Statistical Science, 1990, 5, 463-485.

Ogburn, E. L., Rotnitzky, A., & Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77*(2), 373–396.

Owen, A. B. (2001). *Empirical likelihood*. Chapman & Hall/CRC Press.

Pettitt, A. N. (1976). A two-sample Anderson-Darling rank statistic. *Biometrika, 63*(1), 161–168.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Rubin, D. B. (1986). Statistics and causal inference: Comment: which ifs have causal answers. *Journal of the American Statistical Association, 81*(396), 961–962.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine, 17*(3), 371–385.

Scholz, F. W., & Stephens, M. A. (1987). K-sample Anderson-Darling tests. *Journal of the American Statistical Association, 82*(399), 918–924.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association, 101*(476), 1607–1618.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica, 21*(1), 5–42.